ScaDS.AI
DRESDEN LEIPZIG
CENTER FOR SCALABLE DATA ANALYTICS AND ARTIFICIAL INTELLIGENCE

Connecting the Right Dots: Entity Resolution on Knowledge Graphs

Presented @ ScaDS.AI Summer School 2022

Daniel Obraczka

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

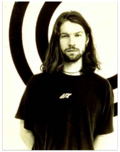
SACHSEN
Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

Users have complex information needs

”In what year did Richard David James
win a Grammy?”

Users have complex information needs

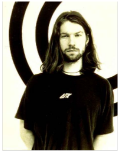
"In what year did Richard David James win a Grammy?"



ScaDS.AI DRESDEN LEIPZIG
TECHNISCHE UNIVERSITÄT DRESDEN
UNIVERSITÄT LEIPZIG

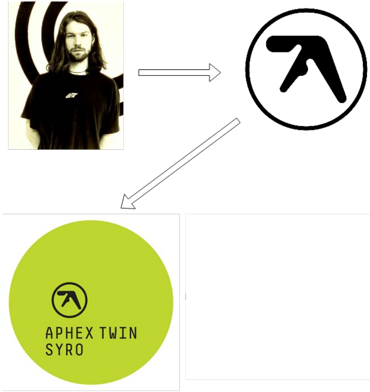
Users have complex information needs

"In what year did Richard David James win a Grammy?"



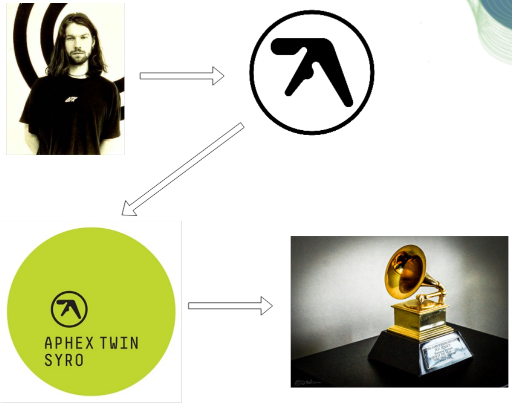
Users have complex information needs

"In what year did Richard David James win a Grammy?"



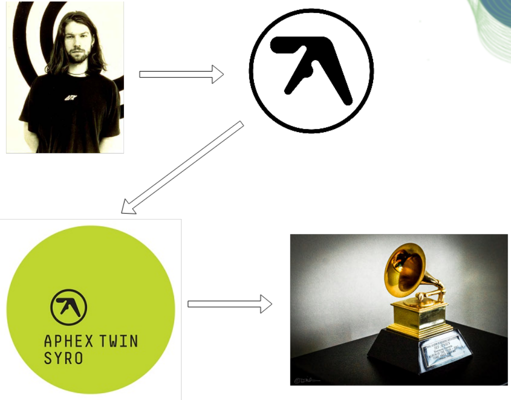
Users have complex information needs

"In what year did Richard David James win a Grammy?"



Users have complex information needs

”In what year did Richard David James win a Grammy?” $\Rightarrow$ 2015



E-commerce example

E-commerce marketplaces have to detect identical products from different shops

Specifications

Acer Aspire E1-572-34014G50Mnkk (Aspire E1 Series)

Processor: Intel Core i3-4010U 2 x 1.7 GHz, Haswell

Graphics adapter: Intel HD Graphics 4400

Display: 15.60 inch 16:9, 1366 x 768 pixel, glossy: no

Weight: 2.2 kg (= 77.6 oz / 4.85 pounds) (= 0 oz / 0 pounds)

Price: 500 Euro

Item#: N82E16834314429

Acer Laptop Aspire E E1-572-6459 Intel Core i3 4th Gen 4010U (1.7GHz) 4GB Memory 500GB HDD Intel HD Graphics 4400 15.6" Windows 7 Home Premium 64-Bit

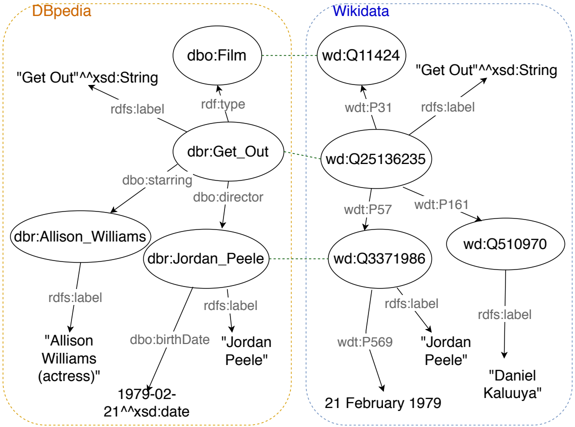
Be the first to review this product...

Check more best sellers of "Laptops / Notebooks"

Check more deals of "Laptops / Notebooks"

Check more lowest price of "Laptops / Notebooks"

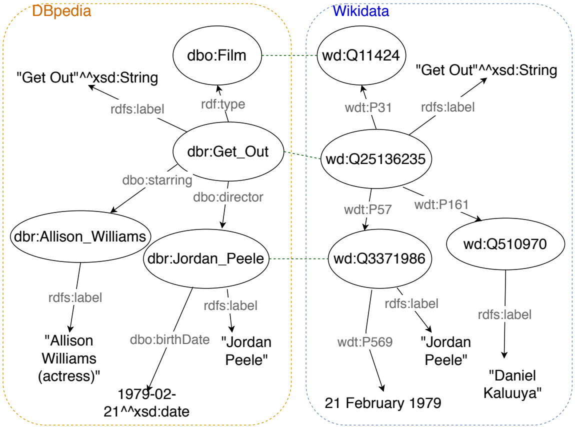
KGs pose specific problems



Flexible schema (usually) means:

- Many entity types
- different (number of) attributes
- various relationship types

KGs pose specific problems



Flexible schema (usually) means:

- Many entity types
- different (number of) attributes
- various relationship types

⇒ Challenging for classical entity resolution systems

Overview

1. Introduction to Entity Resolution on Knowledge Graphs
2. Knowledge Graph Embedding-based approaches
3. Problems with KGE-based approaches

Overview

1. Introduction to Entity Resolution on Knowledge Graphs
2. Knowledge Graph Embedding-based approaches
3. Problems with KGE-based approaches

Disclaimer: Not comprehensive, only overview!

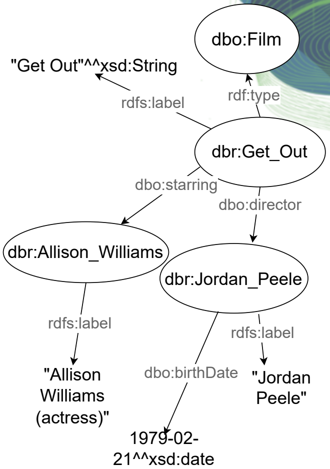
Entity Resolution on Knowledge Graphs

A KG is a tuple $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T})$ where:

- $\mathcal{E}$ is the set of entities
- $\mathcal{R}$ is the set of relation predicates
- $\mathcal{A}$ is the set of attribute predicates
- $\mathcal{V}$ is the set of attribute values
- $\mathcal{T}$ is the set of triples

relation triple: (h, r, t) with $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$

attribute triple: (e, a, v) with $e \in \mathcal{E}$, $a \in \mathcal{A}$ and $v \in \mathcal{V}$



ScaDS.AI DRESDEN LEIPZIG
TECHNISCHE UNIVERSITÄT DRESDEN
UNIVERSITÄT LEIPZIG

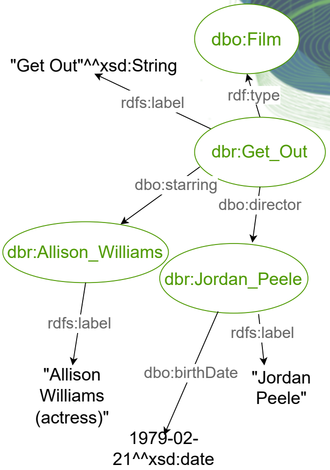
Entity Resolution on Knowledge Graphs

A KG is a tuple $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T})$ where:

- $\mathcal{E}$ is the set of entities
- $\mathcal{R}$ is the set of relation predicates
- $\mathcal{A}$ is the set of attribute predicates
- $\mathcal{V}$ is the set of attribute values
- $\mathcal{T}$ is the set of triples

relation triple: (h, r, t) with $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$

attribute triple: (e, a, v) with $e \in \mathcal{E}$, $a \in \mathcal{A}$ and $v \in \mathcal{V}$



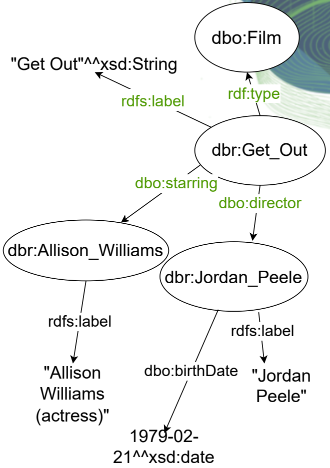
Entity Resolution on Knowledge Graphs

A KG is a tuple $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T})$ where:

- $\mathcal{E}$ is the set of entities
- $\mathcal{R}$ is the set of relation predicates
- $\mathcal{A}$ is the set of attribute predicates
- $\mathcal{V}$ is the set of attribute values
- $\mathcal{T}$ is the set of triples

relation triple: (h, r, t) with $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$

attribute triple: (e, a, v) with $e \in \mathcal{E}$, $a \in \mathcal{A}$ and $v \in \mathcal{V}$



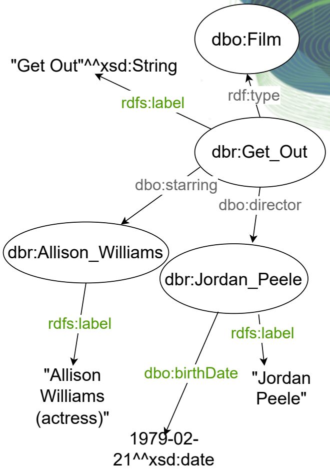
Entity Resolution on Knowledge Graphs

A KG is a tuple $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T})$ where:

- $\mathcal{E}$ is the set of entities
- $\mathcal{R}$ is the set of relation predicates
- $\mathcal{A}$ is the set of attribute predicates
- $\mathcal{V}$ is the set of attribute values
- $\mathcal{T}$ is the set of triples

relation triple: (h, r, t) with $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$

attribute triple: (e, a, v) with $e \in \mathcal{E}$, $a \in \mathcal{A}$ and $v \in \mathcal{V}$



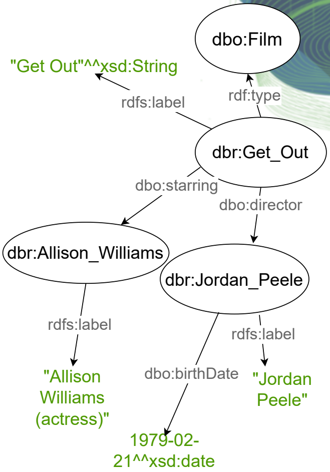
Entity Resolution on Knowledge Graphs

A KG is a tuple $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T})$ where:

- $\mathcal{E}$ is the set of entities
- $\mathcal{R}$ is the set of relation predicates
- $\mathcal{A}$ is the set of attribute predicates
- $\mathcal{V}$ is the set of attribute values
- $\mathcal{T}$ is the set of triples

relation triple: (h, r, t) with $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$

attribute triple: (e, a, v) with $e \in \mathcal{E}$, $a \in \mathcal{A}$ and $v \in \mathcal{V}$



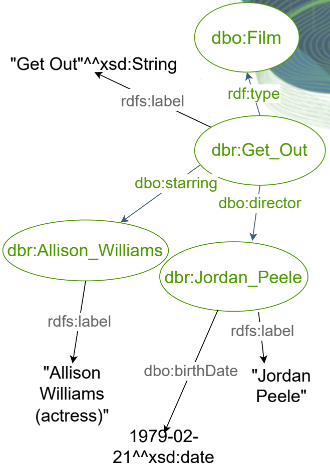
Entity Resolution on Knowledge Graphs

A KG is a tuple $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T})$ where:

- $\mathcal{E}$ is the set of entities
- $\mathcal{R}$ is the set of relation predicates
- $\mathcal{A}$ is the set of attribute predicates
- $\mathcal{V}$ is the set of attribute values
- $\mathcal{T}$ is the set of triples

relation triple: (h, r, t) with $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$

attribute triple: (e, a, v) with $e \in \mathcal{E}$, $a \in \mathcal{A}$ and $v \in \mathcal{V}$



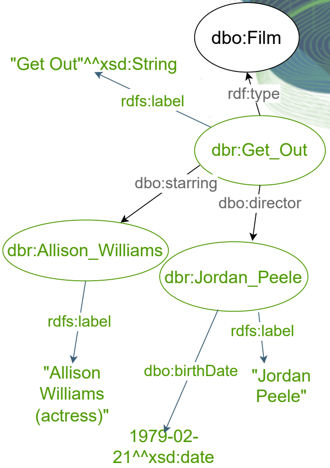
Entity Resolution on Knowledge Graphs

A KG is a tuple $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{A}, \mathcal{V}, \mathcal{T})$ where:

- $\mathcal{E}$ is the set of entities
- $\mathcal{R}$ is the set of relation predicates
- $\mathcal{A}$ is the set of attribute predicates
- $\mathcal{V}$ is the set of attribute values
- $\mathcal{T}$ is the set of triples

relation triple: (h, r, t) with $h, t \in \mathcal{E}$ and $r \in \mathcal{R}$

attribute triple: (e, a, v) with $e \in \mathcal{E}$, $a \in \mathcal{A}$ and $v \in \mathcal{V}$



Entity Resolution on Knowledge Graphs

Definition

Task: Given graphs $\mathcal{G}_1, \mathcal{G}_2$ find mapping $\mathcal{M} = \{(e_1, e_2) \in \mathcal{E}_1, \mathcal{E}_2 | e_1 \equiv e_2\}$, where $\equiv$ refers to the equivalence relation

Variations:

- Clean-Clean: both sources are duplicate-free
- Clean-Dirty: one source is duplicate-free
- Dirty-Dirty: no source is duplicate-free
- Multi-source
- Incremental: Continuously integrate new data without full recomputation

Entity Resolution on Knowledge Graphs (Challenges)

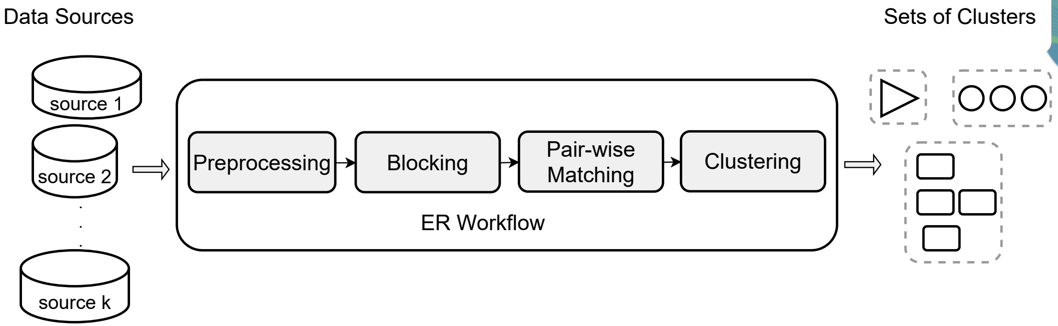
- **Volume**: KGs can be huge (e.g. 10^8 entities in Wikidata)
- **Variety**: KGs usually have heterogeneous schemata
- **Velocity**: KGs are usually updated continuously, necessitating ER solutions, that can tackle this aspect

Entity Resolution on Knowledge Graphs (Challenges)

- **Volume**: KGs can be huge (e.g. 10^8 entities in Wikidata)
- **Variety**: KGs usually have heterogeneous schemata
- **Velocity**: KGs are usually updated continuously, necessitating ER solutions, that can tackle this aspect

Many systems focus on one (or more) of these aspects, but there is no one-size fits all system

General ER workflow



Blocking

- ER complexity is quadratic a-priori (have to compare all entities with each other)

first name	last name
Johan	Müller
Alicia	Mühler
Sara	Fink

first name	last name
Johannes	Müller
Alicia	Mühler
Sara	Fink

Blocking

- ER complexity is quadratic a-priori (have to compare all entities with each other)
- Blocking avoids unnecessary matches by e.g. only comparing entities with same first character in specific attribute

first name	last name
Johan	Müller
Alicia	Mühler
Sara	Fink

first name	last name
Johannes	Müller
Alicia	Mühler
Sara	Fink

Blocking

- ER complexity is quadratic a-priori (have to compare all entities with each other)
- Blocking avoids unnecessary matches by e.g. only comparing entities with same first character in specific attribute

Plethora of approaches exist, for an overview see Papadakis et al., “Blocking and Filtering Techniques for Entity Resolution: A Survey”, 2020

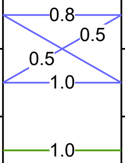
first name	last name
Johan	Müller
Alicia	Mühler
Sara	Fink

first name	last name
Johannes	Müller
Alicia	Mühler
Sara	Fink

Matching

- Based an attribute similarities create a similarity graph
- Many different similarity functions exist (e.g. edit-distance, soundex, etc.)
- (Supervised) machine learning approaches can be used to learn match probabilities

first name	last name
Johan	Müller
Alicia	Mühler
Sara	Fink



first name	last name
Johannes	Müller
Alicia	Mühler
Sara	Fink

Clustering

- Given a similarity graph find clusters of matching entities
- Different clustering strategies perform well based on setting
- For binary clean-clean matching: Hungarian algorithm[1]
- For multi-source clean-clean: CLIP[2]

---

[1] *Jonker and Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems", 1987*

[2] *Saeedi, Peukert, and Rahm, "Using Link Features for Entity Clustering in Knowledge Graphs", 2018*

FAMER

Fast **M**ulti-Source **E**ntity **R**esolution

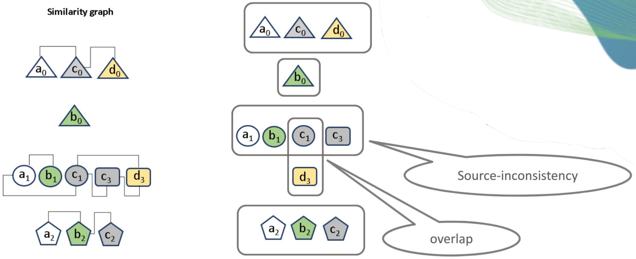
- Build on Apache Flink
- Provides a variety of Blocking methods
- Configurable similarity measures for pairwise matching
- Several clustering algorithms to find matching entities

Saeedi et al., "Scalable Matching and Clustering of Entities with FAMER", 2018

FAMER's CLIP Clustering

Produces
- Source consistent clusters
- No overlap

Prioritize links based on
- Link strength
- Strong, Normal, Weak
- Link degree
- Similarity value

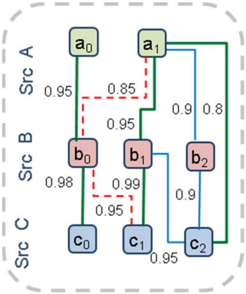


FAMER's CLIP Clustering

Produces
- Source consistent clusters
- No overlap

Prioritize links based on
- Link strength
- Strong, Normal, Weak
- Link degree
- Similarity value

– Link Strength
- **Strong**
- **Normal**
- **Weak**



Some Other Selected Entity Resolution Tools

- DeepMatcher[3]: Uses deep learning with various different techniques to aggregate pre-trained word embeddings of attributes
- LIMES[4]: Relies on triangle equality to avoid blocking while still preventing unnecessary comparisons
- JedAI[5]: Build on Spark, provides schema-agnostic blocking schemes which can also be applied to RDF data
- WInte.r[6]: Modular framework enabling the integration of multiple (web) data sources

---

[3] *Mudgal et al., "Deep learning for entity matching: A design space exploration", 2018*

[4] *Ngomo et al., "LIMES: A Framework for Link Discovery on the Semantic Web", 2021*

[5] *Papadakis et al., "JedAI[3] : beyond batch, blocking-based Entity Resolution", 2020*

[6] *Lehmberg, Bizer, and Brinkmann, "WInte.r - A Web Data Integration Framework", 2017*

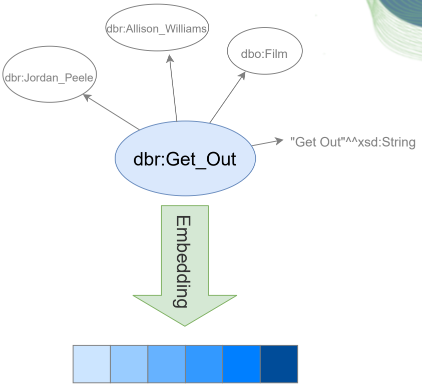
Entity Alignment with Knowledge Graph Embeddings

Knowledge Graph Embeddings (KGEs)

Transform entities into a dense vector

If successful:

- similar entities close in the embedding space
- relational information retained

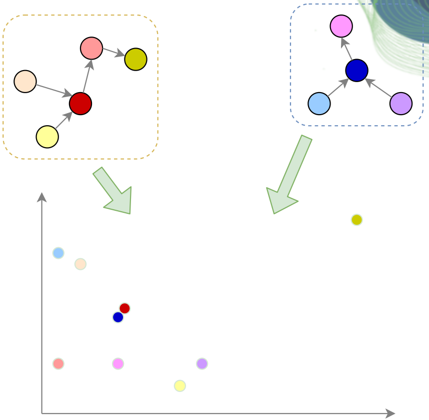


Knowledge Graph Embeddings (KGEs)

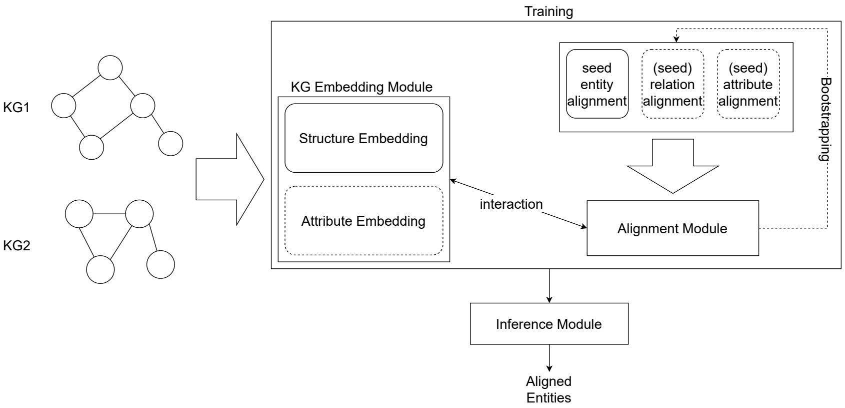
Transform entities into a dense vector

If successful:

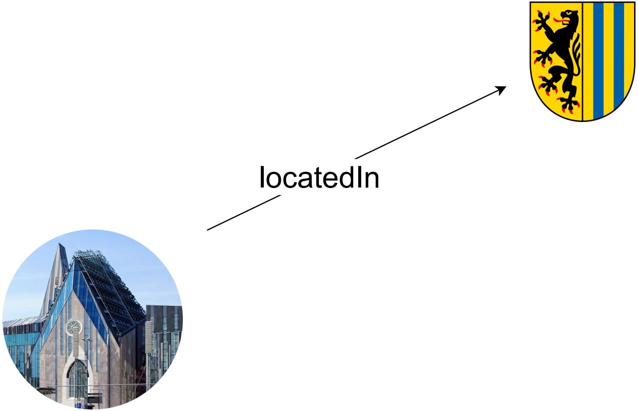
- similar entities close in the embedding space
- relational information retained



Entity Alignment with KGEs Overview

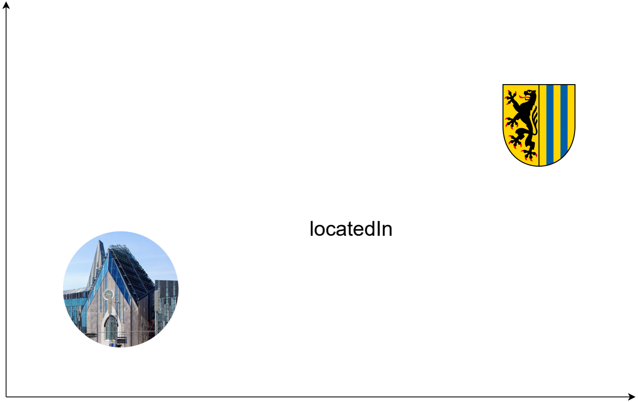


Structure embedding: translation-based



Bordes et al., "Translating embeddings for modeling multi-relational data", 2013

Structure embedding: translation-based

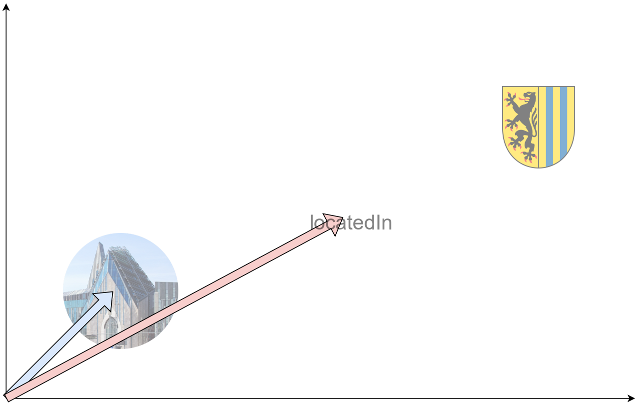


TransE model:

- For triple (h, r, t) minimize $f(h, r, t) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||$

Bordes et al., "Translating embeddings for modeling multi-relational data", 2013

Structure embedding: translation-based

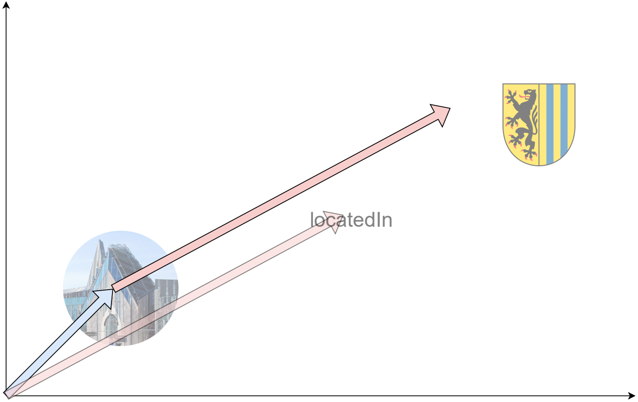


TransE model:

- For triple (h, r, t) minimize $f(h, r, t) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||$

Bordes et al., "Translating embeddings for modeling multi-relational data", 2013

Structure embedding: translation-based

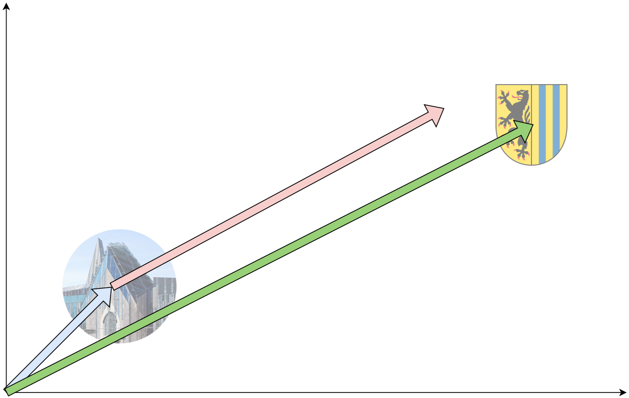


TransE model:

- For triple (h, r, t) minimize $f(h, r, t) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||$

Bordes et al., "Translating embeddings for modeling multi-relational data", 2013

Structure embedding: translation-based

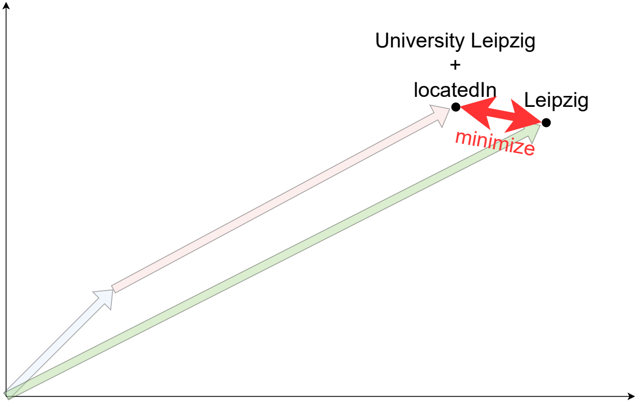


TransE model:

- For triple (h, r, t) minimize $f(h,r,t) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||$

Bordes et al., "Translating embeddings for modeling multi-relational data", 2013

Structure embedding: translation-based

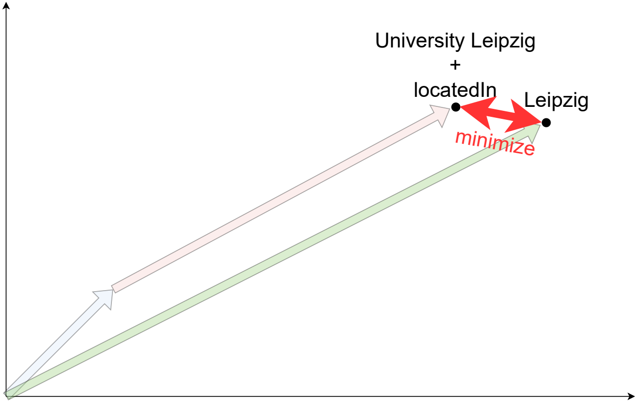


TransE model:

- For triple (h, r, t) minimize $f(h, r, t) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||$

Bordes et al., "Translating embeddings for modeling multi-relational data", 2013

Structure embedding: translation-based



TransE model:

- For triple (h, r, t) minimize $f(h, r, t) = ||\mathbf{h} + \mathbf{r} - \mathbf{t}||$
- This function scores the plausibility of a triple (true triples should have value of 0)
- Corrupted triples (for which either h or t is replaced) should score high

Bordes et al., "Translating embeddings for modeling multi-relational data", 2013

Translation-based

- Simple translational model incapable of modelling one-to-many relationships
- Many extensions: e.g. TransR[7] uses relation-specific spaces
- Kazemi and Poole[8] show that translational models operating in euclidean spaces are severely limited in types relations they can learn
- This shortcoming is for example addressed by HyperKG[9], which operates in the hyperbolic space and is more expressive than previous translational models

---

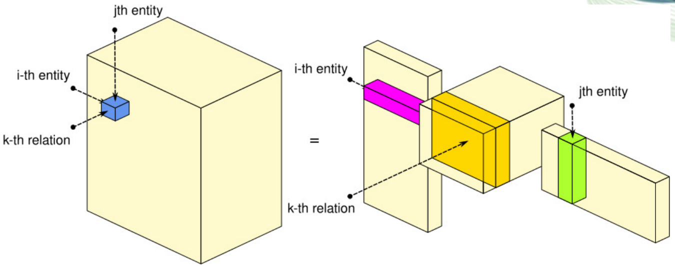
[7] *Lin et al., "Learning Entity and Relation Embeddings for Knowledge Graph Completion", 2015*

[8] *Kazemi and Poole, "Simple embedding for link prediction in knowledge graphs", 2018*

[9] *Kolyvakis, Kalousis, and Kiritsis, "Hyperbolic Knowledge Graph Embeddings for Knowledge Base Completion", 2020*

Tensor-factorization

- KG as 3-order tensor
- Score plausibility of triple (h, r, t) as $f(h, r, t) = \mathbf{h}^T\mathbf{W}\mathbf{t}$



Graphic of RESCAL taken from Maximilian Nickel's page
Nickel, Tresp, and Kriegel, "A three-way model for collective learning on multi-relational data", 2011

Tensor-factorization

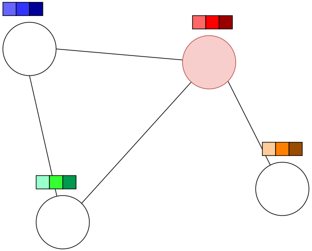
- RESCAL's representation of relations as matrices is costly
- DistMult[10] restricts the relation matrix to a diagonal matrix (but can only model symmetric relations)
- ComplEx[11] extends DistMult in the complex domain and enables modeling of asymmetric relationships
- SimplE[12] represents each entity with two independent vectors via canonical polyadic decomposition. This model is more efficient than e.g. ComplEx, but fully expressive

[10] *Yang et al., "Embedding entities and relations for learning and inference in knowledge bases", 2015*

[11] *Trouillon et al., "Complex Embeddings for Simple Link Prediction", 2016*

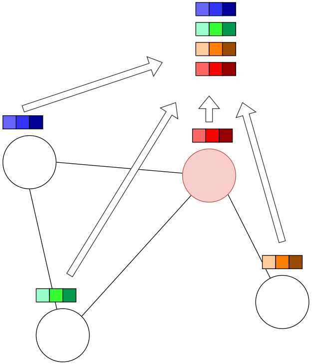
[12] *Kazemi and Poole, "Simple embedding for link prediction in knowledge graphs", 2018*

Graph Convolutional Networks (Intuition)



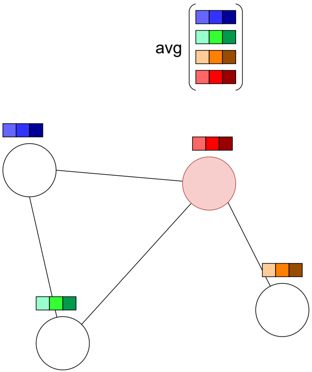
Kipf and Welling, "Semi-Supervised Classification with Graph Convolutional Networks", 2017

Graph Convolutional Networks (Intuition)



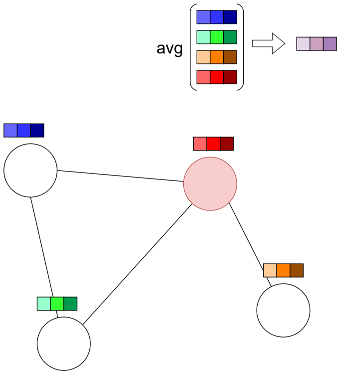
Kipf and Welling, "Semi-Supervised Classification with Graph Convolutional Networks", 2017

Graph Convolutional Networks (Intuition)



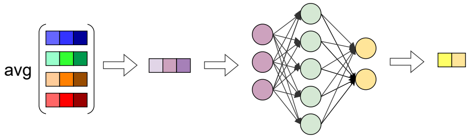
Kipf and Welling, "Semi-Supervised Classification with Graph Convolutional Networks", 2017

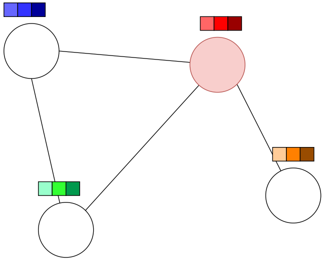
Graph Convolutional Networks (Intuition)



Kipf and Welling, "Semi-Supervised Classification with Graph Convolutional Networks", 2017

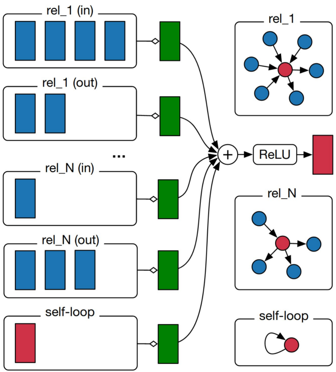
Graph Convolutional Networks (Intuition)





Kipf and Welling, *"Semi-Supervised Classification with Graph Convolutional Networks"*, 2017

Relational Graph Convolutional Networks



- Gather features of neighboring nodes
- Aggregate for each relation type seperately
- Accumulate resulting representation in (normalized) sum
- Send result through activation

Schlichtkrull et al., "Modeling Relational Data with Graph Convolutional Networks", 2018

EA approaches relying only on structure

MTransE[13] uses linear transformation to move entities into same embedding space

- Rely on TransE scoring
- Alignment function measures (dis)similarity between triples of the two graphs:

$$f_{align}(tr_1, tr_2) = ||M_e h_1 - h_2|| + ||M_r r_1 - r_2|| + ||M_e t_1 - t_2||$$

13 *Chen et al., "Multilingual knowledge graph embeddings for cross-lingual knowledge alignment", 2017*

EA approaches relying only on structure

BootEA[14] uses bootstrapping to introduce likely entity matches as training data:

- Given two (likely) matching entities e_1, e_2
- Swap entities in triples with their counterpart and add these new triples to graph
- E.g. for a triple (e_1, r, t) add new triple (e_2, r, t)

Model tries to minimize loss from TransE scoring (including generated triples) and a specific alignment loss based on distance of entity embeddings

[14] *Sun et al., "Bootstrapping entity alignment with knowledge graph embedding", 2018*

EA approaches including attribute information

AttrE[15] introduced the use of attribute values

- Align predicates based on string similarity
- Use scoring function for attribute triples
- For attribute values use either
  - averaged character embedding
  - aggregated character embedding by LSTM
  - aggregated n-gram character embedding (worked best)
- Minimize distance between structure and attribute embedding of an embedding

---

15 *Trisedya, Qi, and Zhang, "Entity Alignment between Knowledge Graphs Using Attribute Embeddings", 2019*

EA approaches including attribute information

MultiKE[16] uses three different views for entity embeddings

- *relation-view*: based on TransE (modified with logistic loss)
- *name-view*: for specific ”name property” a concatenation of pre-trained word/character embeddings is sent through an autoencoder
- *attribute-view*: Use a CNN over attribute-value matrix instantiated with word-embeddings of attribute predicates and their values
- For relation/attribute predicates soft alignment is used to find counterparts across KGs, based on similarity of relation/attribute embeddings (above a certain threshold)
- Similar to BootEA, a triple swapping strategy is used to generate more triples with known matches (or soft aligned)

16 *Zhang et al., “Multi-view knowledge graph embedding for entity alignment”, 2019*

Problems with KGE-based approaches

The hubness phenomenon

- Main focus of KGE research was creation

The hubness phenomenon

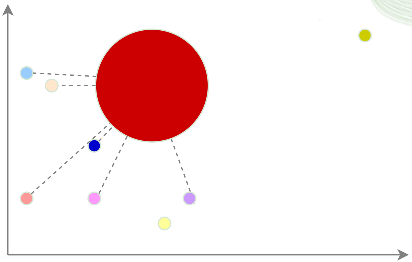
- Main focus of KGE research was creation
- Alignment of KGEs usually relies on Nearest Neighbors

The hubness phenomenon

- Main focus of KGE research was creation
- Alignment of KGEs usually relies on Nearest Neighbors

With increasing dimensionality:

- few points are nearest neighbors (NN) of many points
- many points are NN of no points

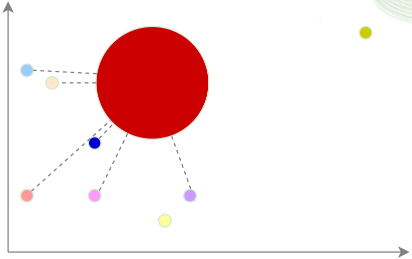


The hubness phenomenon

- Main focus of KGE research was creation
- Alignment of KGEs usually relies on Nearest Neighbors

With increasing dimensionality:

- few points are nearest neighbors (NN) of many points
- many points are NN of no points



$\Rightarrow$ hubness negatively affects alignment quality

Hubness reduction (HR)

Different ideas:

- Centering
- Repair asymmetric relationships

Hubness reduction (HR)

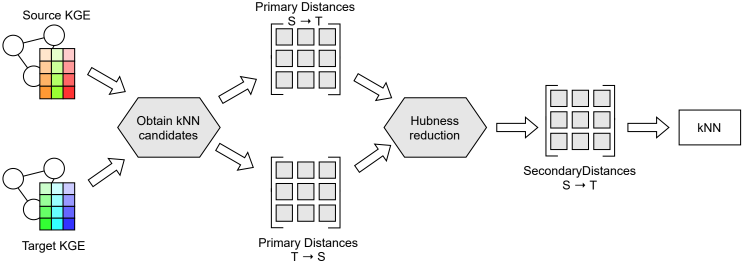
Different ideas:

- Centering
- Repair asymmetric relationships

Overview: Feldbauer and Flexer, “A comprehensive empirical comparison of hubness reduction in high-dimensional spaces”, 2019

kiez

Open-source python library (github.com/dobraczka/kiez)
for hubness-reduced nearest neighbor search
(for entity alignment (with knowledge graph embeddings))



Obraczka and Rahm, "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings", 2021

kiez

Open-source python library (github.com/dobraczka/kiez)
for hubness-reduced nearest neighbor search
(for entity alignment (with knowledge graph embeddings))

Hubness reduction methods:

- Local Scaling *Schnitzer et al., 2012*
- NICDM *Schnitzer et al., 2012*
- CSLS *Lample et al., 2018*
- Mutual Proximity *Schnitzer et al., 2012*
- DisSimLocal *Hara et al., 2016*

kiez

Open-source python library (github.com/dobraczka/kiez)
for hubness-reduced nearest neighbor search
(for entity alignment (with knowledge graph embeddings))

Hubness reduction methods:

- Local Scaling *Schnitzer et al., 2012*
- NICDM *Schnitzer et al., 2012*
- CSLS *Lample et al., 2018*
- Mutual Proximity *Schnitzer et al., 2012*
- DisSimLocal *Hara et al., 2016*

(Approximate) Nearest Neighbor Method:

- Sci-kit learn *Pedregosa et al., 2011*
  - BallTree *Omohundro, 1989*
  - KDTree *Bentley, 1975*
  - Bruteforce
- NMSLIB: HNSW *Malkov, 2018*
- NGT *Iwasaki, 2016*
- Annoy *(github.com/spotify/annoy)*
- Faiss *Johnson, Douze, and Jégou, 2017*

kiez

Open-source python library (github.com/dobraczka/kiez)
for hubness-reduced nearest neighbor search
(for entity alignment (with knowledge graph embeddings))

Hubness reduction methods:

- Local Scaling *Schnitzer et al., 2012*
- **NICDM** *Schnitzer et al., 2012*
- CSLS *Lample et al., 2018*
- Mutual Proximity *Schnitzer et al., 2012*
- DisSimLocal *Hara et al., 2016*

(Approximate) Nearest Neighbor Method:

- Sci-kit learn *Pedregosa et al., 2011*
  - BallTree *Omohundro, 1989*
  - KDTree *Bentley, 1975*
  - Bruteforce
- NMSLIB: HNSW *Malkov, 2018*
- NGT *Iwasaki, 2016*
- Annoy *(github.com/spotify/annoy)*
- Faiss *Johnson, Douze, and Jégou, 2017*

Non-iterative contextual dissimilarity measure

Schnitzer et al., "Local and global scaling reduce hubs in space", 2012

$$NICDM(d_{x,y}) = \frac{d_{x,y}}{\sqrt{\mu_x \mu_y}}$$

Non-iterative contextual dissimilarity measure

Schnitzer et al., “Local and global scaling reduce hubs in space”, 2012

$$NICDM(d_{x,y}) = \frac{d_{x,y}}{\sqrt{\mu_x \mu_y}}$$

mean distance to the k-nearest neighbors

Experiment setup

- 16 aligment tasks:
  - KG samples from DBpedia, Wikidata, YAGO
  - different densities, sizes and even cross-lingual settings

Sun et al., “A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs”, 2020

Experiment setup

- 16 aligment tasks:
  - KG samples from DBpedia, Wikidata, YAGO
  - different densities, sizes and even cross-lingual settings

  Sun et al., “A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs”, 2020

- 15 KG embedding approaches

Experiment setup

- 16 aligment tasks:
  - KG samples from DBpedia, Wikidata, YAGO
  - different densities, sizes and even cross-lingual settings

  Sun et al., “A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs”, 2020

- 15 KG embedding approaches

⇒ 240 KGE pairs

Evaluation Metric

hits@k:

- suited for kNN-based tasks
- counts proportion of true matches in kNN

We use k=50, because we retrieve 50 nearest neighbors

Evaluation Metric

hits@k:

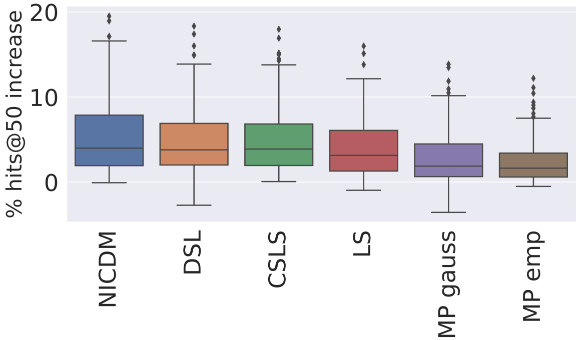
- suited for kNN-based tasks
- counts proportion of true matches in kNN

We use k=50, because we retrieve 50 nearest neighbors

Because absolute hits@k value is largely determined by KGE approach:

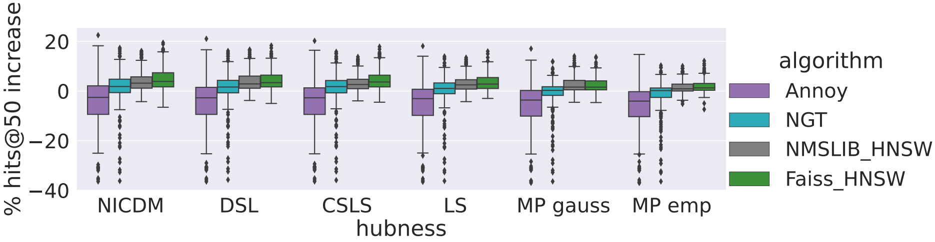
- look at improvement
- compare against no HR with same KGE

Hubness reduction (with exact NN) improves alignment



Improvement in hits@50 compared to no hubness reduction.
Aggregated over KGE approaches and datasets.

Hubness reduction with ANN can improve alignment



Improvement in hits@50 compared to baseline (no HR with exact NN).

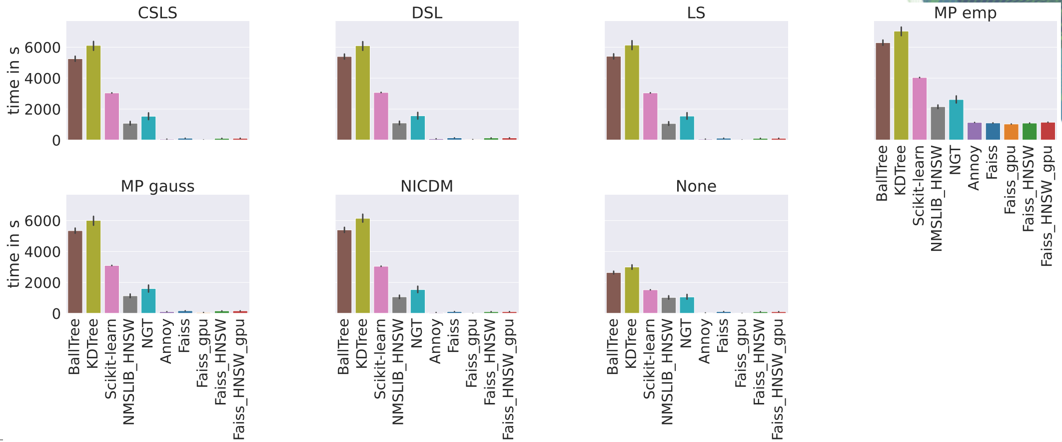
Aggregated over KGE approaches and datasets.

ScaDS.AI DRESDEN LEIPZIG

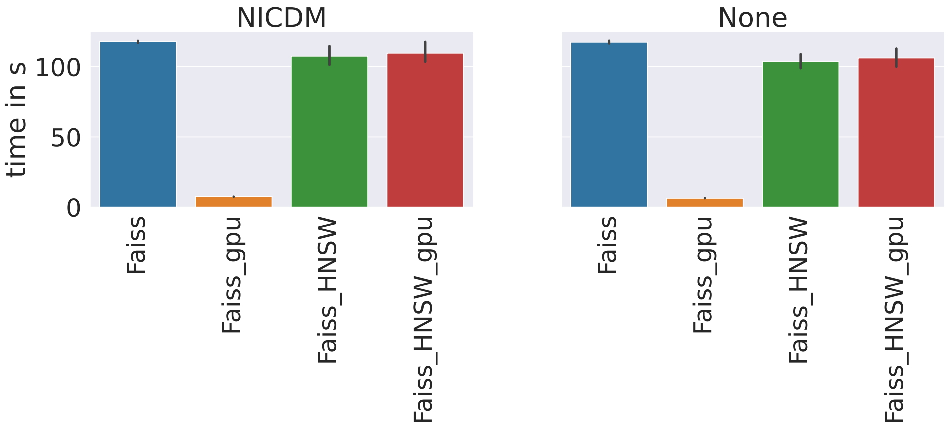
TECHNISCHE UNIVERSITÄT DRESDEN

UNIVERSITÄT LEIPZIG

Speed comparison (large datasets)



Speed comparison (large datasets) only Faiss



Hubness-reduced nearest neighbor search

Verdict:

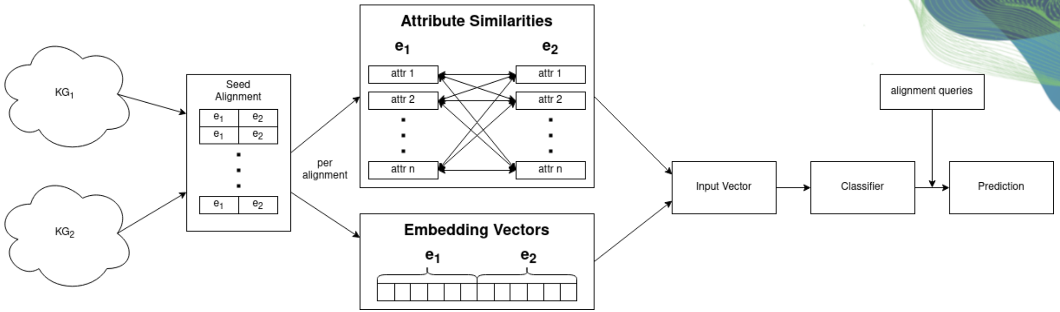
- Hubness reduction improves alignment results
- Using Faiss with NICDM gives improvements at virtually no cost w.r.t speed
- For larger datasets Faiss's HNSW implementation can be used

⇒ Hubness reduction largely offsets decrease in alignment quality when using *approximate* nearest neighbor algorithm while still retaining speed advantage

How to give more prominence to attribute values?

- KGE-based approaches heavily emphasize graph structure
- There is usually no direct attribute similarity calculated between entities

EAGER: Embedding-assisted Entity Resolution for KGs



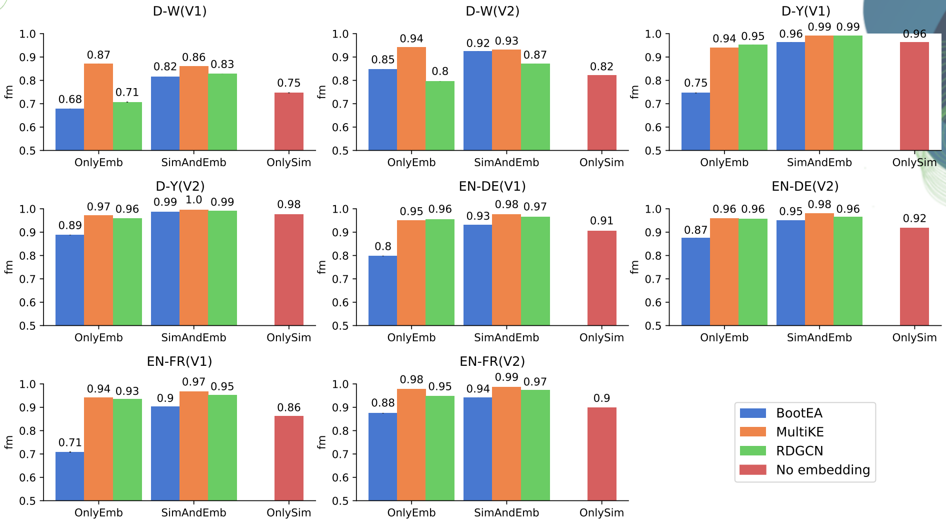
Obraczka, Schuchart, and Rahm, "Embedding-Assisted Entity Resolution for Knowledge Graphs", 2021

Experiment Setup

Investigate performance of combination trough ablation study

→Three different inputs for EAGER:

- **OnlyEmb**: Only use embeddings
- **OnlySim**: Only use attribute similarities
- **SimAndEmb**: Use both



Results for 100K datasets (using MLP as classifier)

More problems with KGE-based approaches

For a critical look:

Leone et al., “A Critical Re-evaluation of Neural Methods for Entity Alignment”, 2022

- Most approaches are evaluated on datasets with (unrealistic) 1-to-1 assumption
- Approaches are costly and have problems scaling
- Currently no way of handling unseen entities
- Authors adapted PARIS[17] to incorporate seed alignment and could generally outperform SOTA KGE-based methods

---

[17] *Suchanek, Abiteboul, and Senellart, “PARIS: Probabilistic Alignment of Relations, Instances, and Schema”, 2011*

Future directions for KGE-based methods

- Combinations of KGE-based methods and techniques from record linkage can be fruitful (see results from EAGER or other work[a])
- Usage of KGE-based methods as blocking strategy has not been explored yet
- Many benchmark datasets consist mostly of ”easy” matches, use-cases with low lexical similarity across matches might be where KGE-based methods shine
- Making KGE-based methods more scalable is a must
- Unsupervised KGE-based methods are still rare

---

[a] *Qi et al., “Unsupervised Knowledge Graph Alignment by Probabilistic Reasoning and Semantic Embedding”, 2021*

What did we learn?

- Data integration has been a long studied field and KGs pose specific challenges (especially volume & variety)
- ”Classical” ER tools rely mostly on attribute similarity for match decisions
- Basic intuition behind KGEs was presented
- KGE-based methods rely mostly on graph structure and incorporate attribute information via pre-trained word embeddings
- KGE-based methods have still much room for improvement, but combining ”old” and new methods might be a fruitful future direction

What did we learn?

- Data integration has been a long studied field and KGs pose specific challenges (especially volume & variety)
- ”Classical” ER tools rely mostly on attribute similarity for match decisions
- Basic intuition behind KGEs was presented
- KGE-based methods rely mostly on graph structure and incorporate attribute information via pre-trained word embeddings
- KGE-based methods have still much room for improvement, but combining ”old” and new methods might be a fruitful future direction

Contact:

obraczka@informatik.uni-leipzig.de

github.com/dobraczka

dobraczka

Thank you for your attention!

- Bentley, Jon Louis (1975). “Multidimensional Binary Search Trees Used for Associative Searching”. In: *Commun. ACM* 18.9, pp. 509–517. DOI: 10.1145/361002.361007. URL: http://doi.acm.org/10.1145/361002.361007.
- Bordes, Antoine et al. (2013). “Translating embeddings for modeling multi-relational data”. In.
- Chen, Muhao et al. (2017). “Multilingual knowledge graph embeddings for cross-lingual knowledge alignment”. In: *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1511–1517. ISSN: 10450823. DOI: 10.24963/ijcai.2017/209.
- Feldbauer, Roman and Arthur Flexer (2019). “A comprehensive empirical comparison of hubness reduction in high-dimensional spaces”. In: *Knowledge and Information Systems* 59.1, pp. 137–166. ISSN: 02193116. DOI: 10.1007/s10115-018-1205-y. URL: https://doi.org/10.1007/s10115-018-1205-y.
- Hara, Kazuo et al. (2016). “Flattening the density gradient for eliminating spatial centrality to reduce hubness”. In: *30th AAAI Conference on Artificial Intelligence, AAAI 2016*.

- Iwasaki, Masajiro (2016). "Pruned Bi-directed K-nearest neighbor graph for proximity search". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9939 LNCS, pp. 20–33. ISBN: 9783319467580. DOI: 10.1007/978-3-319-46759-7_2.
- Johnson, Jeff, Matthijs Douze, and Hervé Jégou (2017). "Billion-scale similarity search with GPUs". In: *arXiv preprint arXiv:1702.08734*.
- Jonker, Roy and A. Volgenant (1987). "A shortest augmenting path algorithm for dense and sparse linear assignment problems". In: *Computing* 38.4, pp. 325–340. DOI: 10.1007/BF02278710. URL: https://doi.org/10.1007/BF02278710.
- Kazemi, Seyed Mehran and David Poole (2018). "Simple embedding for link prediction in knowledge graphs". In: *Advances in Neural Information Processing Systems* 2018-Decem (Nips), pp. 4284–4295. ISSN: 10495258.
- Kipf, Thomas N. and Max Welling (2017). "Semi-Supervised Classification with Graph Convolutional Networks". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=SJU4ayYgl.

Kolyvakis, Prodromos, Alexandros Kalousis, and Dimitris Kiritsis (2020). “Hyperbolic Knowledge Graph Embeddings for Knowledge Base Completion”. In: *The Semantic Web - 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31-June 4, 2020, Proceedings*. Ed. by Andreas Harth et al. Vol. 12123. Lecture Notes in Computer Science. Springer, pp. 199–214. DOI: 10.1007/978-3-030-49461-2_12. URL: https://doi.org/10.1007/978-3-030-49461-2_12.

Lample, Guillaume et al. (2018). “Word translation without parallel data”. In: *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*.

Lehmberg, Oliver, Christian Bizer, and Alexander Brinkmann (2017). “WInte.r - A Web Data Integration Framework”. In: *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017*. Ed. by Nadeschda Nikitina et al. Vol. 1963. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-1963/paper506.pdf.

Leone, Manuel et al. (2022). “A Critical Re-evaluation of Neural Methods for Entity Alignment”. In: *Proc. VLDB Endow.* 15.8, pp. 1712–1725. URL: https://www.vldb.org/pvldb/vol15/p1712-arora.pdf.

- Lin, Yankai et al. (2015). "Learning Entity and Relation Embeddings for Knowledge Graph Completion". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. Ed. by Blai Bonet and Sven Koenig. AAAI Press, pp. 2181–2187. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9571.
- Malkov, Yu. A. (2018). "Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 31–33. URL: https://github.com/nmslib/nmslib{\%}0Ahttp://ann-benchmarks.com/hnsw(nmslib).html.
- Mudgal, Sidharth et al. (2018). "Deep learning for entity matching: A design space exploration". In: *Proceedings of the 2018 International Conference on Management of Data*, pp. 19–34.
- Ngomo, Axel-Cyrille Ngonga et al. (2021). "LIMES: A Framework for Link Discovery on the Semantic Web". In: *Künstliche Intell.* 35.3, pp. 413–423. DOI: 10.1007/s13218-021-00713-x. URL: https://doi.org/10.1007/s13218-021-00713-x.
- Nickel, Maximilian, Volker Tresp, and Hans Peter Kriegel (2011). "A three-way model for collective learning on multi-relational data". In: *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 809–816.

- Obraczka, Daniel and Erhard Rahm (2021). "An Evaluation of Hubness Reduction Methods for Entity Alignment with Knowledge Graph Embeddings". In: *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2021, Volume 2: KEOD, Online Streaming, October 25-27, 2021*. Ed. by David Aveiro, Jan L. G. Dietz, and Joaquim Filipe. SCITEPRESS, pp. 28–39. DOI: 10.5220/0010646400003064. URL: https://doi.org/10.5220/0010646400003064.
- Obraczka, Daniel, Jonathan Schuchart, and Erhard Rahm (2021). "Embedding-Assisted Entity Resolution for Knowledge Graphs". In: *Proceedings of the 2nd International Workshop on Knowledge Graph Construction co-located with 18th Extended Semantic Web Conference (ESWC 2021), Online, June 6, 2021*. Ed. by David Chaves-Fraga et al. Vol. 2873. CEUR Workshop Proceedings. CEUR-WS.org. URL: http://ceur-ws.org/Vol-2873/paper8.pdf.
- Omohundro, Stephen M. (1989). *Five Balltree Construction Algorithms*. Tech. rep. International Computer Science Institute.
- Papadakis, George et al. (2020a). "Blocking and Filtering Techniques for Entity Resolution: A Survey". In: *ACM Comput. Surv.* 53.2, 31:1–31:42. DOI: 10.1145/3377455. URL: https://doi.org/10.1145/3377455.

Papadakis, George et al. (2020b). “JedAI^3 : beyond batch, blocking-based Entity Resolution”. In: *Proceedings of the 23rd International Conference on Extending Database Technology, EDBT 2020, Copenhagen, Denmark, March 30 - April 02, 2020*. Ed. by Angela Bonifati et al. OpenProceedings.org, pp. 603–606. DOI: 10.5441/002/edbt.2020.74. URL: https://doi.org/10.5441/002/edbt.2020.74.

Pedregosa, F. et al. (2011). “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Qi, Zhiyuan et al. (2021). “Unsupervised Knowledge Graph Alignment by Probabilistic Reasoning and Semantic Embedding”. In: *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*. Ed. by Zhi-Hua Zhou. ijcai.org, pp. 2019–2025. DOI: 10.24963/ijcai.2021/278. URL: https://doi.org/10.24963/ijcai.2021/278.

Saeedi, Alieh, Eric Peukert, and Erhard Rahm (2018). “Using Link Features for Entity Clustering in Knowledge Graphs”. In: *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*. Ed. by Aldo Gangemi et al. Vol. 10843. Lecture Notes in Computer Science. Springer, pp. 576–592. DOI: 10.1007/978-3-319-93417-4_37. URL: https://doi.org/10.1007/978-3-319-93417-4_37.

- Saeedi, Alieh et al. (2018). “Scalable Matching and Clustering of Entities with FAMER”. In: *Complex Syst. Informatics Model. Q.* 16, pp. 61–83. DOI: 10.7250/csimq.2018-16.04. URL: https://doi.org/10.7250/csimq.2018-16.04.
- Schlichtkrull, Michael Sejr et al. (2018). “Modeling Relational Data with Graph Convolutional Networks”. In: *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*. Ed. by Aldo Gangemi et al. Vol. 10843. Lecture Notes in Computer Science. Springer, pp. 593–607. DOI: 10.1007/978-3-319-93417-4_38. URL: https://doi.org/10.1007/978-3-319-93417-4_38.
- Schnitzer, Dominik et al. (2012). “Local and global scaling reduce hubs in space”. In: *Journal of Machine Learning Research* 13. ISSN: 15324435.
- Suchanek, Fabian M., Serge Abiteboul, and Pierre Senellart (2011). “PARIS: Probabilistic Alignment of Relations, Instances, and Schema”. In: *Proc. VLDB Endow.* 5.3, pp. 157–168. DOI: 10.14778/2078331.2078332. URL: http://www.vldb.org/pvldb/vol5/p157_fabianmsuchanek_vldb2012.pdf.
- Sun, Zequn et al. (2018). “Bootstrapping entity alignment with knowledge graph embedding”. In: *IJCAI International Joint Conference on Artificial Intelligence* 2018-July, pp. 4396–4402. ISSN: 10450823. DOI: 10.24963/ijcai.2018/611.

- Sun, Zequn et al. (2020). "A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs". In: *Proc. VLDB Endow.* 13.11, pp. 2326–2340. URL: http://www.vldb.org/pvldb/vol13/p2326-sun.pdf.
- Trisedya, Bayu Distiawan, Jianzhong Qi, and Rui Zhang (2019). "Entity Alignment between Knowledge Graphs Using Attribute Embeddings". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33, pp. 297–304. ISSN: 2159-5399. DOI: 10.1609/aaai.v33i01.3301297.
- Trouillon, Théo et al. (2016). "Complex Embeddings for Simple Link Prediction". In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*. Ed. by Maria-Florina Balcan and Kilian Q. Weinberger. Vol. 48. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 2071–2080. URL: http://proceedings.mlr.press/v48/trouillon16.html.
- Yang, Bishan et al. (2015). "Embedding entities and relations for learning and inference in knowledge bases". In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–13.
- Zhang, Qingheng et al. (June 2019). "Multi-view knowledge graph embedding for entity alignment". In: vol. 2019-Augus. International Joint Conferences on Artificial Intelligence, pp. 5429–5435. ISBN: 9780999241141. DOI: 10.24963/ijcai.2019/754. URL: http://arxiv.org/abs/1906.02390.